DOCUMENT RESUME

ED 099 430 95 TM 004 308

AUTHOR Eash, Maurice J.; And Others
TITLE Evaluation Designs for Practitioners. TM Report No. 35.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ETS-TM-35
PUB DATE Dec 74
CONTRACT OEC-0-70-3797-519
NOTE 6p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Decision Making; Design; *Education; Educational Research; *Evaluation; Guides; *Program Evaluation; *Research Design

ABSTRACT

Practitioners are not afforded the luxury of ideal laboratory conditions. The natural settings of the classroom, the school, or the school system place constraints on the type of data obtainable; hence, educators must work with less than an ideal experimental design. Four evaluation designs used in natural settings are described. Each involves an evaluation study that takes into account a variety of constraints, but nevertheless provides a basis for subsequent program and/or organizational decision. The study includes a true experimental design in a field setting, a nonequivalent control groups design, a time series design, and a no comparison groups design. (Author/RC)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 35 DECEMBER 1974

EVALUATION DESIGNS FOR PRACTITIONERS

Maurice J. Eash
Harriet Talmage
Herbert J. Walberg

Planning and implementing any facet of the educational program call for decision making whether the project concerns the program of an entire school system or the day-to-day practice of a teacher in a single classroom. The interactive nature of the educational process produces a dynamic environment; hence, decisions made at one point in time require reassessment at the next point in time before another round of decisions can begin. Evaluation provides a framework for building a systematic data base to aid in making decisions in school and classroom practice. With an appropriate data base, problems can be reformulated, both potential and actual consequences can be analyzed, and, as a result, the processes can be redirected.

Practitioners are not afforded the luxury of ideal laboratory conditions. The natural settings of the classroom, the school, or the school system place constraints upon the type of data obtainable; hence, educators must work with less than an ideal experimental design.[1]

Four evaluation designs used in natural settings are described in the following sections.[2] Each involves an evaluation study that takes into account a variety of constraints, but nevertheless provides a basis for subsequent program and or organizational decisions. The studies range from a true experimental design, one that necessitates the random assignment of students to experimental and control groups, to a design that lacks both randomization and comparative groups.

In each section, the basic paradigm of the evaluation design is symbolically presented. Four symbols identify the elements of the paradigms: R—randomization; X—treatment; O—observation; and in some cases, DA—design analysis. Subscripts denote specific treatments and observations. Observations (O) to the left of the treatment (X) denote pretest data, and to the right, posttest data. The experimental group symbols appear above the control group symbols. A broken line between the groups indicates nonequivalent groups.

[1] The studies used to illustrate the designs were conducted by the Office of Evaluation Research, College of Education, University of Illinois at Chicago Circle.

[2] For additional designs, the reader may wish to consult Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally Company, 1963.

A True Experimental Design in a Field Setting

R	X	O
R		O

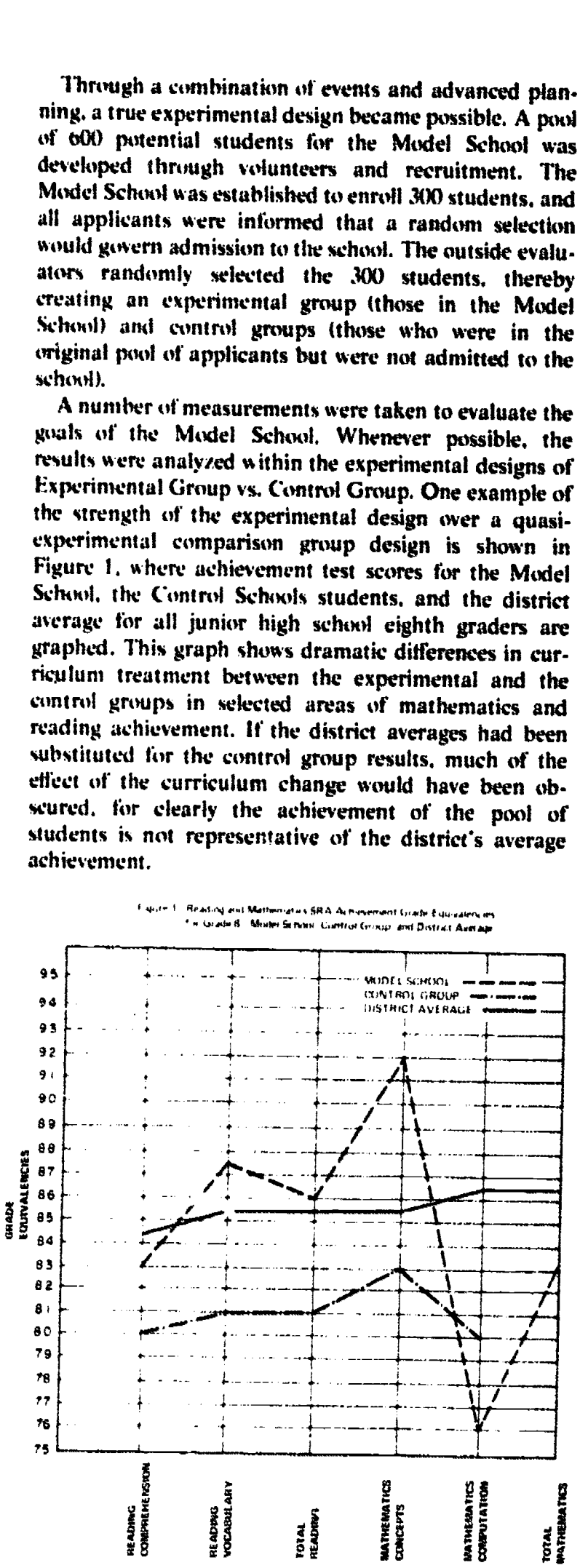
A true experimental design is characterized by its randomization of subjects to treatment—"randomly dividing the litter among treatments"—and is the conventional laboratory-science way of exercising this control. The strength of the design, randomization for control of error, is also a major source of difficulty in field evaluations because studies are conducted where scheduling, teacher preferences in assignment, luncheon arrangements, and a myriad of other considerations enter in. Thus one finds the experimental design infrequently used in reported evaluations. However, because of its power to bring forth more valid findings, we suggest that evaluators search for ways to employ it in field situations. An example drawn from an evaluation of a curriculum model set up under a Title III grant illustrates the power of a true experimental design to bare true differences and the weaknesses of nonrandom comparison groups. Clocktown, a fast growing suburb in a major metropolitan area, received a three-year grant to design a middle school curriculum which would break sharply with the conventional curriculums in the seven other junior high schools. The new curriculum included: 1) greater parent involvement, 2) a more humanistic orientation, 3) promoting greater achievement, 4) promoting more affective growth, 5) integrating pupil personnel services within the curriculum, and 6) offering these changes at a per-pupil cost competitive with the costs in the other junior high schools. After one year of planning, the two-year experimental school opened.

This publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

Through a combination of events and advanced planning, a true experimental design became possible. A pool of 600 potential students for the Model School was developed through volunteers and recruitment. The Model School was established to enroll 300 students, and all applicants were informed that a random selection would govern admission to the school. The outside evaluators randomly selected the 300 students, thereby creating an experimental group (those in the Model School) and control groups (those who were in the original pool of applicants but were not admitted to the school).

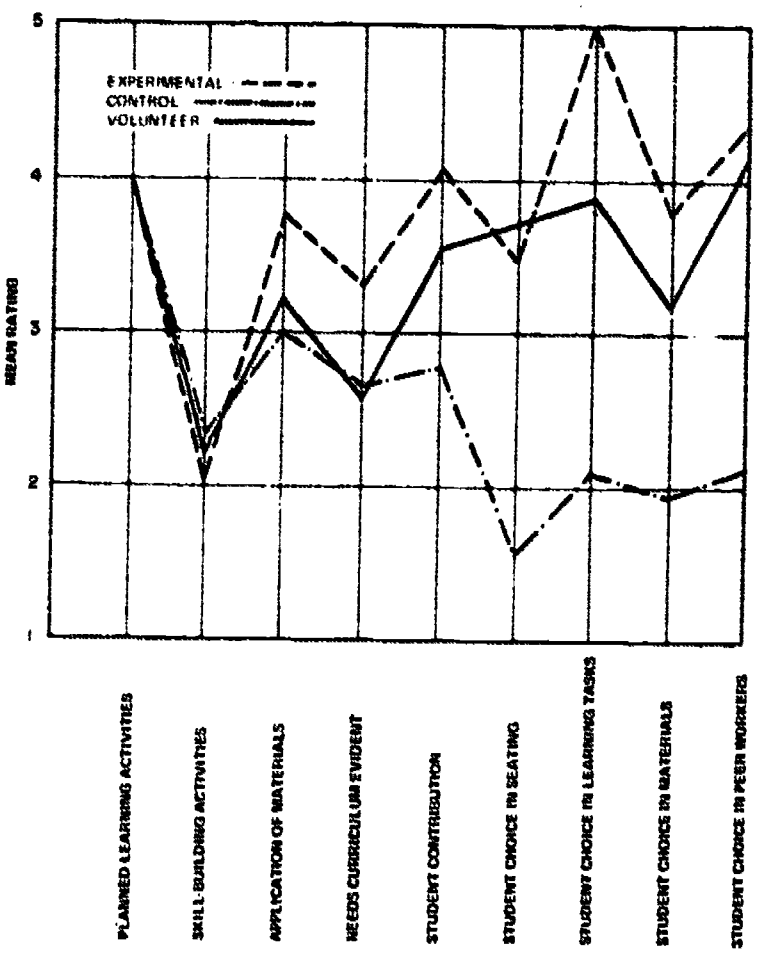
A number of measurements were taken to evaluate the goals of the Model School. Whenever possible, the results were analyzed within the experimental designs of Experimental Group vs. Control Group. One example of the strength of the experimental design over a quasi-experimental comparison group design is shown in Figure 1, where achievement test scores for the Model School, the Control Schools students, and the district average for all junior high school eighth graders are graphed. This graph shows dramatic differences in curriculum treatment between the experimental and the control groups in selected areas of mathematics and reading achievement. If the district averages had been substituted for the control group results, much of the effect of the curriculum change would have been obscured, for clearly the achievement of the pool of students is not representative of the district's average achievement.

Figure 1 Reading and Mathematics SRA Achievement Grade Equivalencies for Grade 8 Model School, Control Group, and District Average



A second example of how true differences are masked is seen when a volunteer group instead of a randomly selected control group is used in a comparison of classroom observations made in volunteer teachers' classrooms. In year two, the control group of classrooms to be observed was randomly selected to obtain a more representative sample of classroom practice to compare with the Model School. The differences are much sharper since the first year volunteer control group classrooms were much closer to the experimental group in practice than were the typical district junior high school classrooms. (See Figure 2.) The experimental design is invaluable to control error and to trace the attribution of results to treatment more clearly. Every effort should be made to use it when the question of curriculum effects is at issue or a summative evaluation is at stake.

Figure 2 Mean Ratings for Randomly Selected Experimental Classrooms, Control Classrooms, and Volunteer Classrooms



Nonequivalent Control-Group Design

O	X	O
O	X	O

It is usually difficult to assign students randomly to classrooms receiving special treatment or to assign teachers randomly within schools to special programs. In the first instance, parents tend to resist changes that vary from the established curriculum without their approval. In the second instance, teachers assigned to new programs involuntarily may affect the outcomes negatively. Through a nonequivalent control-group design, the

handicap due to the lack of randomization is compensated for in several ways.

The Textville School District study concerned the problems of evaluating four new reading series to select one for system-wide adoption. Instructional materials play a significant role in the educational process for 75 percent of the instructional time in the classroom, and 90 percent of the homework time is devoted to these materials. Thus, adoption cannot be taken lightly. Selecting a reading series frequently entails ideological confrontation to the neglect of facts. Publishers display their materials with attractive illustrations and slick copy, and groups of teachers espouse one approach to reading instruction or another as the final solution to all reading problems. Therefore, an evaluation design was developed to serve two purposes: 1) to overcome the difficulties of nonrandomization and 2) to establish a data base for making selection decisions on the basis of facts rather than ideological quibbling.

In designing the evaluation study, the drawbacks of nonrandom assignment of students and teachers to experimental and control groups were taken into consideration by obtaining pretest and posttest data, employing multiple treatments for comparisons with the traditional treatment and comparisons among the treatments, and using the class rather than the individual student as the unit of study. An adaptation of the nonequivalent control-group design is illustrated in Figure 3. Pretest (O_{preT}) and posttest (O_{postT}) reading achievement data were obtained. Data on teacher characteristics (O_1) were initially collected. Subsequent to the introduction of the treatment (X), data were obtained on learning environment variables (O_2: competitiveness, cohesiveness, difficulty, friction, and satisfaction) and on instructional characteristics (O_3: locus of instructional decisions, variety and utilization of materials, and student behaviors).

The Textville schools and teachers were encouraged to participate in the study. Sixty classes from 12 schools were chosen and represented the range of ability, of socioeconomic, racial, and ethnic backgrounds, and of geographic locations found in the district. Assignment to

Figure 3. Nonequivalent Control-Group Design Paradigm

O_{preT}	O_1	X_1	O_{postT}	O_2	O_3
O_{preT}	O_1	X_2	O_{postT}	O_2	O_3
O_{preT}	O_1	X_3	O_{postT}	O_2	O_3
O_{preT}	O_1	X_4	O_{postT}	O_2	O_3
O_{preT}	O_1	$X_{control}$	O_{postT}	O_2	O_3

Figure 4. Assignment to Treatment Matrix

GRADE LEVEL	SERIES X_1	SERIES X_2	SERIES X_3	SERIES X_4	SERIES $X_{control}$	TOTAL
1ST	3	3	3	3	3	15
2ND	3	3	3	3	3	15
3RD	3	3	3	3	3	15
6TH	3	3	3	3	3	15
TOTAL CLASSES	12	12	12	12	12	60

a reading series by grade level is shown in Figure 4. For each reading series, the materials were field-tested in three different schools in grades 1, 2, 3, and 6. In all, the data included 12 different classes per series.

Two constraints were imposed on the design: 1) All four classes in a school field-testing the reading materials must use the same series; and 2) the best educational interest of the students must supersede the design of the study. And, indeed, this came to pass: One class found too many difficulties with the series at the peril of impeding their reading progress, and the class was removed from the study.

The data were analyzed to provide information on four questions:

- Do the classes using one series obtain higher reading scores on the reading achievement posttest than classes using another series?
- Do the classes using one reading series perceive their learning environment differently than do classes using another reading series? Do the learning environment and reading series taken together affect achievement?
- Do selected teacher characteristics in conjunction with a given series affect reading achievement?
- Does instruction differ in classes using different reading series?

Statistical analyses indicated that the pretest score is the single most significant predictor of reading achievement despite teacher characteristics and regardless of the reading series. After the effects of the pretest scores are removed, competitiveness is the only other variable that predicts reading achievement. The higher the competitiveness in the learning environment, the lower the reading achievement. There are no significant correlations between competitiveness and reading series, teacher characteristics, or instructional characteristics.

The final selection decision for the Textville School District shifted away from an emphasis on ideological issues such as phonics-oriented vs. nonphonics-oriented reading approaches or linguistic vs. nonlinguistic

reading approaches. In place of these, attention was focused on the instructional aspects of a reading program that tend to reduce competitiveness, and on such concerns as the district's philosophy of reading, cost factors, implementation problems, and the degree of teacher dependence on outside support.

Time-Series Design

19?2	19?3		19?3	19?4
O	X	O	O	X	O

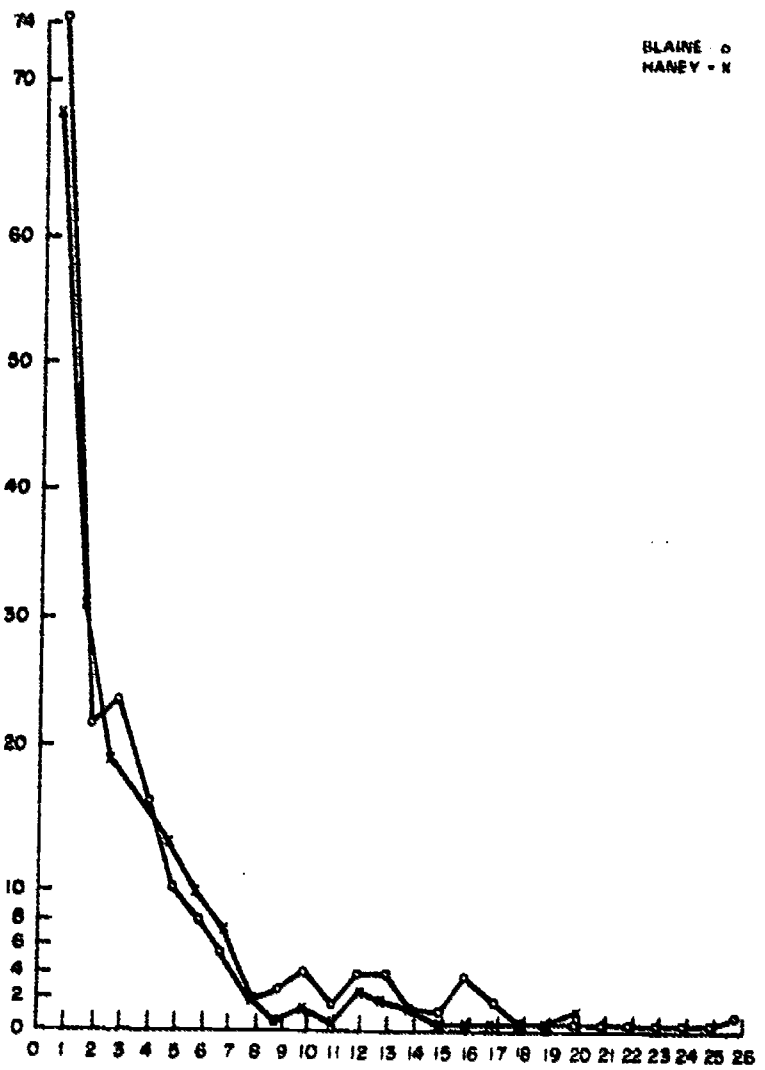
Practitioners are frequently faced with the necessity of making major program changes which reorganize curriculum and structural arrangements. Not infrequently, such changes are precipitated by external forces that are impatient with the setting up of an evaluation design that would require the establishing of control groups before the change is made. In these cases, data are frequently desperately needed by administrative decision makers if they are not to be at the mercy of rumor and pressure groups. Such was the case of the Parkland School District, which was suddenly under a legal mandate to integrate its schools. *De facto* segregation resulting from segregated housing placed practically all the black population in one elementary school and the white population in six schools, and produced segregation up through grade 6. The junior high schools were integrated in name, but not always in reality, for the students segregated themselves by race in the lunchroom and on the playground. Faced with a legal mandate to bus students to achieve equal racial proportions in all seven elementary schools, Parkland administrators requested an outside evaluator to help them set up an evaluation design that would provide basic data on these questions: 1) What effects does the structural reorganization required by busing have on student achievement and on the learning environment? 2) What data would be useful for program planning and for alerting the administration to potential difficulties?

The evaluation was hampered by the inability to set up control groups through randomization. Moreover, since the entire school system was involved, no separate control groups were available. Within these limitations, it was decided to use a time-series design for a two-year period that would allow within-the-group comparisons, use a multiple collection of data, and give a reading on several indices of progress. Experience indicates that over the two years many productive hypotheses were generated and an invaluable data base for charting progress in achievement, race relations, and classroom instruction was established.

A pretest and posttest on general achievement was given every child in the fall and spring. Since there were previous local norms available, these data quieted fears that integration was destroying achievement. A learning environment measure, administered in the spring, revealed that further curriculum planning was needed to improve the learning environment for both white and black students in different schools. An analysis of the learning environment and achievement measures revealed that some schools appeared to be much more successful than others in providing a stimulating learning environment and promoting achievement. While the lack of adequate controls limited generalizations or conclusions, these data did pinpoint areas for closer investigation by administrators and teachers. One of the more immediately useful applications of evaluation data came when rumors of the deterioration of discipline in one school swept the community. The recently administered learning environment inventory profile calmed both the school board and the public by its demonstration that the students in this school perceived their environment very much as did their counterparts in other schools, and that there was no greater conflict or disruption in their school than in the others.

A third area of data was an analysis of the records of disciplinary cases in the junior high schools. These again provided some short-term data as the basis of decisions, since the offenses that took up most of administrators' and guidance counselors' time were being committed by a very small group of students. (See Figure 5.)

Figure 5. Summary Graph of Frequencies of Behavioral Incidents



Interracial problems were not as prevalent as intraracial problems. A second year of charting these behavioral incidents showed that the concentration of social services on the few major offenders had removed them from the behavioral records in the second year. In addition, it was found that interracial conflict had decreased. Thus, one is led to conclude that the time-series design provides a useful data base for decision making in a situation where tensions induced by structural changes cry for the voice of rationality. One must admit that these data have limited generalizability, but they have been invaluable in the context in which they are collected and in demonstrating that evaluation can serve several purposes in applied settings.

No Comparison Group Design

Design Analysis	X — O X — O X — O	Design Analysis

Not infrequently, an evaluator is confronted with a program that is to be used but is being undertaken with restraints that forestall the use of control groups. Is usefulness of evaluation forestalled under these circumstances, and must one retreat to the rhetoric of castigating shortsightedness in the developer? The fourth example deals with such a problem.

An outside private agency provided funds to increase and improve the teaching of the arts in schools. Launched from very broad objectives, "to enable parents and community leaders to use the arts as communication tools," the agency requested evaluation assistance to improve the series of workshops that it had designed for teachers.

From the workshops' guides that were presented and from the funding proposal, an analysis of workshop activities to achieve the goals was prepared. The activities proved to be a better source of goals than the diffuse general objectives. The evaluation design was concerned with: 1) Were the activities being taught in the workshops? 2) Were they perceived as useful by teachers since they incorporated creative and nonconventional teaching approaches? 3) Were they being implemented in classrooms and did they maintain the integrity of the activities?

The evaluators were not permitted to gather data from control classes in the schools, nor were they to observe the instructors assigned to the workshops. The design of the evaluation structured the gathering of data by analyzing the program and developing an activity analysis, which was then converted into an instrument to be used by teachers to evaluate workshop activities on four dimensions: 1) the workshop participants' reaction, 2) whether teachers used any one activity in the classroom, 3) the students' reactions, and 4) ease of implementation. A second source of data was gathered from a pretest and a posttest of learning environments in the workshop participants' classrooms. A third source of data was observation in classrooms where teachers taught the workshop activities to their students. A fourth source of data was a standardized teacher evaluation questionnaire to evaluate the workshops.

From these data an analysis was made of the workshops, and recommendations were rendered on which workshops and what activities were most useful in the classroom. As this evaluation progressed, feedback sessions were held with workshop directors to assist them in conducting the next semester's workshops. Evaluation in this case focused on providing clari[illegible] to a group of program developers who were worki[illegible] nbiguous area. Although many of the traditional [illegible]an evaluation design are lacking, these are c[illegible]t be generated and comparisons that can be made to s[illegible]ape the educational product. In the sense of serving to improve practice through the establishment of a data base and promoting meaningful comparisons for practitioners, the evaluation design remains true to its calling in bringing rationality to play on educational activities.

Cooperative Planning in Evaluation

Evaluation is often viewed by practitioners as being outside their reach: The designs are incomprehensible, the data are too costly to gather, the participants are threatened by the potential of the findings, and the effects, efforts, and efficiency cannot be evaluated with any degree of objectivity anyway. Our experience, gathered over a wide variety of projects, would indicate that practitioners are handicapped by too narrow a view of evaluation and by their failure to systematically build an evaluation design into projects. Moreover, troublesome problems are not approached through an evaluation design which in its use converts rhetoric to a factual base, as was illustrated in the example on the reading series. In short, decision making and choice taking are blind through the lack of evaluation designs which open up options and permit an earlier use of correctives in program planning.

To provide for an evaluation design in the early stages makes for a more open commitment to the major goals of a project, and establishes a degree of latitude for shifting direction based on evidence which often is denied when the program participants' personal commitments to a project deepen with effort. Evaluation can serve to keep the focus on the quest for a better way to provide education as opposed to espousing a dogma of "the way to provide quality education." If evaluation is seen as a necessary part of projects and problem solving, the use of evaluators and evaluation findings becomes as significant as the appropriate use of evaluation designs. Findings must be implemented to be effective in decision making.

At the Office of Evaluation Research, we have found that cementing an early working relationship between

the evaluators and the practitioners is the best guarantee of the use of evaluation findings. As outside evaluators, this entails building an evaluation design early in the project with inputs from practitioners on their needs for data. In another context, we have referred to this process as a *coactional relationship*[3] where the two parties are engaged in a mutual task with a commitment to the discovery of options and the search for truth. Extra effort is required from the evaluators to explain designs and their strengths and constraints; but these early sessions also build the foundation of commitment to follow the findings wherever they may lead. The process is coactional in that the evaluators perceive the context in which the evaluation design is being used and it is early on plans for implementation of findings at appropriate junctures. Our contention is that many evaluation reports are superfluous because they are ill-timed to the schedule of information needs of practitioners, or return findings that are arcane and remote from the decisions that are pressing the decision maker. We see, as imperative to success, the need to be sensitive to the roles of the evaluators and practitioners and their relationships in building evaluation designs. The four evaluation designs described illustrate applications of a methodology in a field context. Brevity did not permit the description of roles and relationships, though they are implicit in the applications. Appropriate use of evaluation designs, we contend, can bring rationality into play in field-based problems and can improve educational practice.

[3]Maurice J. Eash, "Transactional Evaluation of Classroom Practice," in *Studies in Transactional Evaluation*, ed. Robert M. Rippey, Berkeley: McCuchan Pub. Corp., 1973.